Computer Science and Application 计算机科学与应用, 2026, 16(2), 277-288
Published Online February 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.162058

Hans汉斯

U型网络的轻量化设计及端侧部署研究

冯佳祥，刘　洋，赵一丁，黄孟轩，于欣鑫

长春理工大学数学与统计学院，吉林 长春

收稿日期：2026年1月3日；录用日期：2026年2月3日；发布日期：2026年2月11日

摘　要

针对现有医学影像分割模型参数规模大、计算复杂度高而难以在资源受限的边缘设备上高效部署的问题，本文提出了一种基于国产瑞芯微RK3588开发板的轻量化U型网络设计及量化部署方案。研究首先基于U-Net架构引入空洞门控注意力(DGA)、反转外部注意力(IEA)及特征桥接模块，构建了轻量化网络 MALUNet以平衡特征提取能力与计算开销；结合一次性层剪枝与归一化知识蒸馏技术对模型进行深度压缩，并利用rknn-toolkit2完成NPU端侧的量化部署。在ISIC2017数据集上的实验结果显示，优化后的MALUNetGlobalAtt学生模型在保持较高分割精度(mIoU为0.8126)的前提下，单样本推理时间较原始模型降低了96%，验证了该方案在国产边缘计算平台上实现医学影像实时智能分析的可行性与优越性。

关键词

RK3588，MALUNet，NPU，量化部署

Research on Lightweight Design and Edge Deployment of U-Shaped Networks

Jiaxiang Feng, Yang Liu, Yiding Zhao, Mengxuan Huang, Xinxin Yu

School of Mathematics and Statistics, Changchun University of Science and Technology, Changchun Jilin

Received: January 3, 2026; accepted: February 3, 2026; published: February 11, 2026

Abstract

To address the challenges that existing medical image segmentation models suffer from large parameter sizes and high computational complexity, making them difficult to be efficiently deployed on resource-constrained edge devices, this paper proposes a lightweight U-shaped network design and quantized deployment scheme based on the domestic Rockchip RK3588 development board. Firstly, based on the U-Net architecture, the study constructs a lightweight network, MALUNet, by incorporating Dilated Gated Attention (DGA), Inverted External Attention (IEA), and feature bridge blocks to balance feature extraction capability with computational cost. Furthermore, the model is

文章引用: 冯佳祥, 刘洋, 赵一丁, 黄孟轩, 于欣鑫. U 型网络的轻量化设计及端侧部署研究[J]. 计算机科学与应用, 2026, 16(2): 277-288. DOI: 10.12677/csa.2026.162058

deeply compressed by combining one-shot layer pruning and normalized knowledge distillation techniques, and the quantized deployment on the NPU is completed using rknn-toolkit2. Experimental results on the ISIC2017 dataset demonstrate that the optimized MALUNetGlobalAtt student model reduces the single-sample inference time by 96% compared to the original model while maintaining high segmentation accuracy (mIoU of 0.8126). This validates the feasibility and superiority of this scheme for real-time intelligent analysis of medical images on domestic edge computing platforms.

Keywords

RK3588, MALUNet, NPU, Quantized Deployment

1. 引言

医学影像技术作为现代医学诊断与临床决策的重要支撑，在疾病筛查、病灶定位以及治疗方案制定等方面发挥着不可替代的作用。随着医学成像设备分辨率的不断提升以及影像数据规模的持续增长，如何从复杂的医学影像中快速、准确地提取关键结构信息，已成为医学图像分析领域亟需解决的核心问题之一。医学影像图像分割通过对感兴趣区域进行精确划分，为病灶检测、器官测量和病理分析提供了可靠依据，因此在临床辅助诊断中具有重要研究价值和应用前景。

近年来，随着深度学习技术的快速发展，基于卷积神经网络的医学影像分割方法取得了显著进展。其中，U 型网络由于其编码器–解码器对称结构以及跳跃连接机制[1]，能够在有效提取高层语义信息的同时保留图像的细节特征，在医学影像分割任务中表现出较高的分割精度和良好的泛化能力。基于 U 型网络的分割模型已被广泛应用于脑部影像、肺部影像以及肿瘤分割等多个医学场景[2] [3]，并逐渐成为医学影像分割领域的主流方法。然而，此类模型通常具有网络结构复杂、参数规模较大以及计算量高等特点，对计算平台的存储资源和浮点运算能力提出了较高要求[4]。

随着智慧医疗和数字医疗的不断推进，医学影像分析系统逐步从传统的中心化计算模式向更加灵活、高效的边缘计算模式转变[5]。在临床应用场景中，实时性和稳定性往往具有较高要求，而依赖云端服务器进行集中处理不仅会引入额边的通信时延，还可能受到网络带宽和数据隐私等因素的制约。边缘计算通过将数据处理与模型推理任务下沉至靠近数据源的边缘设备，有效缩短了数据传输路径，降低了系统整体时延，同时提升了数据安全性和系统响应能力，为医学影像智能分析提供了新的技术支撑。

边缘计算在医学影像应用中具有明显优势，但受限于边缘设备算力和存储资源的限制[6]，现有医学影像分割模型难以直接部署和高效运行。因此，如何在保证分割精度的前提下，对模型进行结构优化和推理加速，成为医学影像分割算法工程化落地过程中面临的重要挑战。本文基于此以瑞芯微 RK3588 平台作为边缘计算硬件基础[7]，对 U 型网络医学影像分割模型进行轻量化设计和量化部署，实现医学影像分割模型在 RK3588 平台上的高效部署与运行。

2. U 型网络的轻量化设计和量化部署方法

2.1. U 型网络的轻量化设计——MALUNet 网络

MALUNet [8]是基于 U 型网络架构进行轻量化设计后得到的医学影像分割模型。考虑到实际应用中

计算资源的限制，尤其是对于移动医疗设备等低功耗平台，MALUnet 在保持 UNet 核心架构的基础上通过集合四个轻量化的注意力机制模块，实现了模型的高效性和高性能之间的平衡。该模型在设计上注重减少参数量与计算复杂度，同时最大化保留分割精度，尤其是在处理复杂的医学影像数据时。接下来的部分将详细介绍这四个模块的设计理念与实现方法，分别是：空洞门控注意力块(Dilated Gated Attention Block, DGA)、反转外部注意力块(Inverted External Attention Block, IEA)、通道注意力桥接块(Channel Attention Bridge Block, CAB)和空间注意力桥接块(Spatial Attention Bridge Block, SAB)。实验证明 MALUNet 能够在尽可能少的计算资源下，达到较为卓越的医学影像分割效果，为低功耗平台上的医学影像智能处理提供了可行的解决方案。MALUNet 架构图如下图 1 所示：

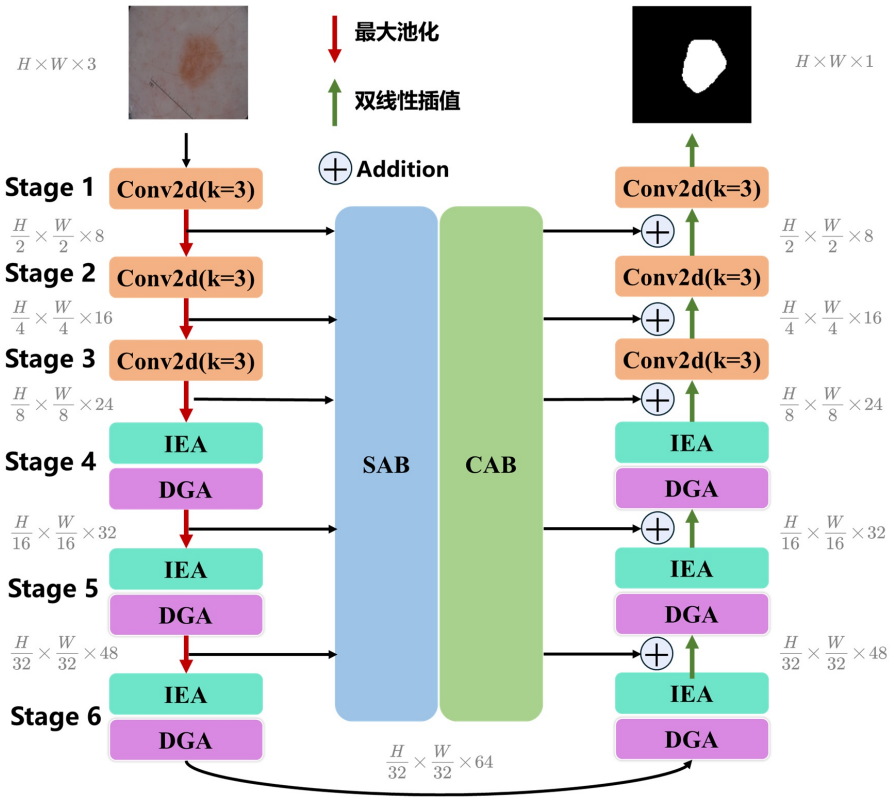

Figure 1. The illustration of MALUNet architecture
图 1. MALUNet 架构示意图

医学影像分割任务属于密集预测问题，因此模型需要具备对整体结构和局部细节信息的理解能力。全局信息有助于模型从整体层面把握目标区域与背景之间的空间关系，而局部信息有助于提高对边界细节和形态特征分割的准确率。

基于此本文引入 DGA 模块[8]，通过融合空洞卷积与门控注意力机制，实现对全局信息与局部信息的特征提取。如图 2(a)，本模块由两个单元组成，分离空洞卷积单元首先利用不同膨胀率的卷积操作对特征图进行处理，使网络能够同时捕获大尺度结构信息和小尺度细节特征。而门控注意力单元对提取到的特征进行自适应调制，使模型能够抑制冗余信息并强化对目标区域的关注。IEA 是在外部注意力机制的基础上引入倒置结构设计的一种改进模块[8] [9]，如图 2(b)，能够实现在保持全局建模能力并进一步降低计算复杂度，增强特征表达的稳定性。如图 2(c)，CAB 用于在通道维度上实现多阶段特征的信息交互与有效融合。CAB 将各阶段特征首先通过全局信息提取操作进行压缩[8]，然后不同阶段的通道特征在通

道维度上进行融合，并通过轻量化的映射生成对应的通道注意力权重。最后使得模型能够识别哪个阶段的拓展通道是重要的。而 SAB 用于在特征融合过程中学习空间位置之间的依赖关系[8]，如图 2(d)，以增强关键区域的表达能力。

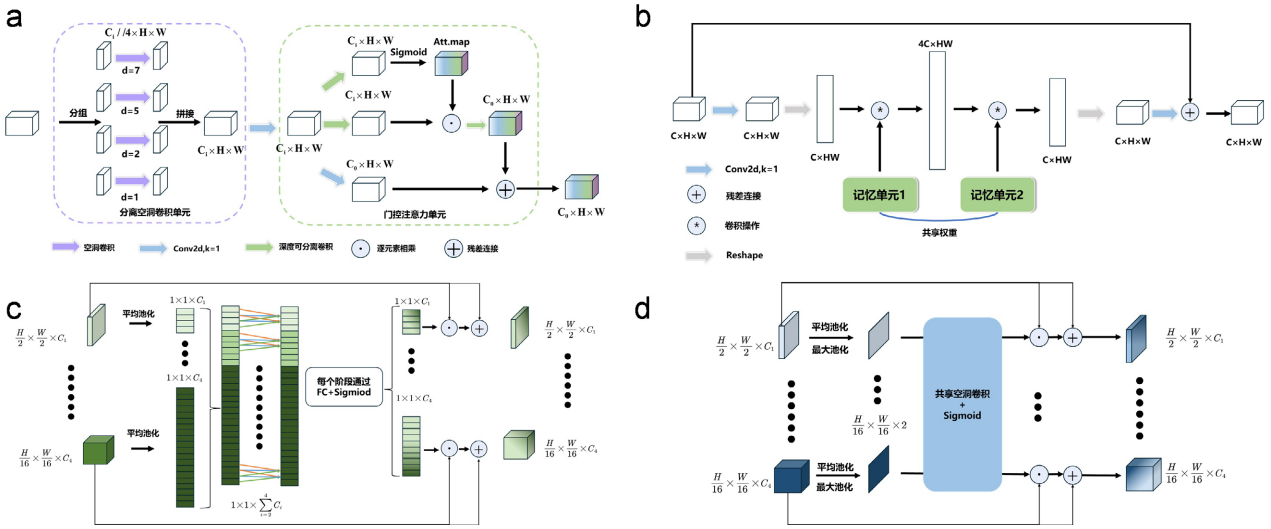

Figure 2. The illustration of the attention module architecture
图 2. 注意力模块架构示意图

2.2. 轻量化部署方法

随着深度神经网络的迅速发展，其模型规模与计算复杂度也随之显著增长。但在边缘设备等资源受限平台上，深度模型实现端侧部署面临着存储开销、推理延迟与能耗问题等问题。基于此，学界逐步形成了以剪枝、蒸馏和量化压缩为代表的三类轻量化部署方法体系。

模型剪枝是最早被系统研究的轻量化手段之一，其基本思想是通过消除网络中的冗余参数或冗余结构，降低模型复杂度。LeCun 等人最早提出基于二阶泰勒展开的 Optimal Brain Damage 方法[10]，通过近似估计单个权重对损失函数的影响程度，从而实现精细化的权重剪枝。Hassibi 与 Stork 在此基础上提出的 Optimal Brain Surgeon 方法[11]考虑到权重之间的相互依赖关系，提高了剪枝精度。Han 等人提出了基于权值幅值的剪枝方法[12]，该方法通过假设小幅值权重对模型输出贡献有限，在训练完成后直接移除低于阈值的权重，并辅以微调过程以恢复性能。近年来，研究重点逐步从非结构化剪枝转向以通道、卷积核或网络层作为剪枝对象的结构化剪枝，可以更好地适配实际硬件架构，本文即将提到的一次性层剪枝就是一种结构化剪枝。

知识蒸馏通过模型间的知识迁移实现轻量化部署。该思想最早由 Hinton 等人提出[13]，其核心在于利用性能较强的教师模型输出的软标签，引导学生模型学习更丰富的类别关系信息，从而在较小模型规模下获得接近教师模型的性能。Romero 等人提出的特征蒸馏方法通过约束中间特征表示实现更深层次的知识传递[14]。

量化压缩则从数值表示的角度降低模型的存储与计算成本，其核心目标是使用低精度数据表示网络参数与中间激活。早期量化研究主要集中在权重量化领域，其核心思想是将原本以浮点数表示的网络权重映射为低比特整数表示。随着深度学习模型规模的不断扩大，Courbariaux 等人提出了二值化与低比特网络[15]，表明在一定条件下低精度表示仍可维持可接受的模型性能。Jacob 等人提出了后训练量化方法[16]，通过在模型训练完成后对权重和激活进行线性映射，实现无需重新训练的快速部署。

2.2.1. 一次性层剪枝

一次性层剪枝的核心思想在于度量每一网络层对整体模型性能的相对重要性[17]，并在满足约束条件的前提下移除冗余层。设原始神经网络由 L 个可剪枝层组成，记第 l 层的参数集合为 W_l，网络整体参数为 $W=\{W_1,W_2,\cdots,W_L\}$，训练目标函数可表示为：

$$L(W)=E_{(x,y)\sim D}\left[\ell\left(f(x;W),y\right)\right], \tag{1}$$

其中 $f(\cdot)$ 表示网络映射，$\ell(\cdot)$ 为损失函数。

一次性层剪枝的关键在于构建层级重要性评分函数 S_l，用于衡量移除第 l 层对损失函数的影响程度。常见做法是基于敏感性分析或近似误差评估，将剪枝问题转化为对损失变化的估计。在忽略高阶项的情况下，移除第 l 层可视为对参数 W_l 施加零化操作，其引起的损失变化可近似表示为：

$$\Delta L_l \approx \left|\frac{\partial L}{\partial W_l}\cdot W_l\right|. \tag{2}$$

ΔL_l 反映了当前参数状态下第 l 层对目标函数的瞬时贡献程度，可作为层重要性的近似度量。基于所有层的重要性评分，一次性层剪枝通常在全局约束条件下完成剪枝决策。在给定参数预算 C 的条件下，剪枝问题将转化为如下优化目标：

$$\min_m \sum_{l=1}^{L} m_l\cdot S_l \quad \text{s.t.} \sum_{l=1}^{L} m_l\cdot c_l \le C, \tag{3}$$

其中 $m=\{m_1,m_2,\cdots,m_L\}$ 为二值掩码变量，$m_l=1$ 表示保留第 l 层，$m_l=0$ 表示剪除该层，c_l 表示对应层的计算或参数开销。通过求解该优化问题，可以在单次决策中确定最终的网络结构。

2.2.2. 归一化蒸馏

归一化蒸馏的核心原理是通过对特征蒸馏损失进行归一化处理[17]，使学生模型能够更加充分地吸收教师模型在不同层次上的表征信息，适用于剪枝后结构发生变化的网络重训练阶段。

设教师模型与学生模型在第 i 个阶段输出的特征表示分别为 f_i^T 与 f_i^S，传统特征蒸馏通常通过最小化二者之间的欧氏距离来实现知识约束，其损失函数可表示为：

$$L_{Feat}=\sum \mathbb{E}\left\|f_i^T-f_i^S\right\|_2^2. \tag{4}$$

上式表示不同阶段特征的数值尺度决定了其在总损失中的相对贡献，当某些特征幅值显著大于其他阶段时，其损失项将主导梯度更新过程，从而削弱整体蒸馏效果。因此，归一化蒸馏通过计算教师模型在各阶段特征的范数信息构造归一化系数，对特征蒸馏损失进行重加权处理。归一化特征蒸馏损失可形式化表示为：

$$L_{Feat}^{norm}=\sum \mathbb{E}\alpha_i\left\|f_i^T-f_i^S\right\|_2^2, \tag{5}$$

其中 α_i 表示第 i 个阶段对应的归一化权重。

在整体训练目标中，归一化蒸馏通常与任务损失及输出蒸馏损失共同作用，从而在性能保持与结构压缩之间实现平衡。通过这种方式，学生模型在学习任务目标的同时，能够在不同层级上稳定地对齐教师模型的中间表示，为剪枝或结构简化后的模型提供更加可靠的性能恢复路径。

2.2.3. 量化压缩

量化压缩的核心在于构建连续浮点空间到离散整数空间之间的映射关系。设某一层的浮点权重为 w，其量化表示为 $\hat{w}$，常见的线性量化过程可表示为[18]：

$$\hat{w} = clip\left(\left[\frac{w}{s}\right] + z\right) \tag{6}$$

其中 s 表示量化尺度因子，用于控制浮点值到整数值的映射比例，z 表示零点偏移，用于保证零值在量化空间中的精确表示。

在推理阶段，量化模型以整数形式执行乘加运算，由于整数运算在 RK3588 平台有更高的吞吐效率，量化压缩能够减少模型体积并降低推理延迟。在量化感知训练场景下，为保证梯度的可传播性，量化函数通常采用直通估计方式对不可导算子进行近似，使模型参数能够在低精度约束下稳定收敛。

3. RK3588 开发板编译流程

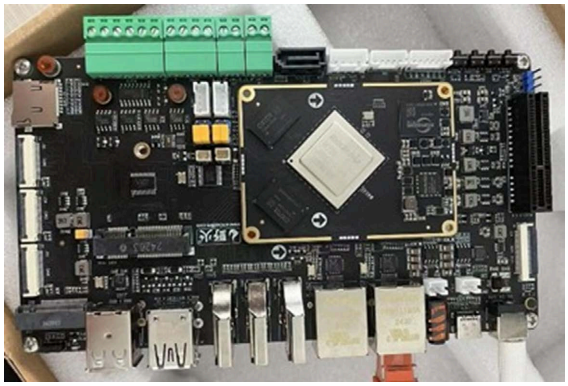

Figure 3. Physical view of the RK3588
图 3. RK3588 实物图

随着深度学习技术对边缘计算需求的日益增长，国产芯片产业持续发展，打破了长期的技术壁垒。瑞芯微电子(Rockchip)推出的新一代旗舰级芯片 RK3588 展现出了卓越的性能表现。RK3588 内置了针对深度学习加速的三核 NPU，能够提供 6TOPS 的混合精度算力支持，具备强大的通用计算能力。因此，RK3588 很适合应用于对实时性与精度要求严苛的医学影像分割任务，为深度学习算法在端侧部署场景提供了硬件基础。RK3588 开发板实物图见图 3。

3.1. ONNX 模型

ONNX 模型是一种针对机器学习模型设计的开放式文件格式标准，它定义了一套通用的算子集与数据流图规范。本项目中 ONNX 模型为连接训练环境与端侧部署环境的中间介质，旨在将基于特定框架训练得到的医学影像分割模型转化为一种硬件无关的标准表达形式。

3.2. RKNN 模型

RKNN 模型作为瑞芯微平台专属的深度学习推理格式，是实现算法在 RK3588 芯片 NPU 上高效运行的核心载体。利用配套的 RKNN-Toolkit2 工具链，我们将上一阶段的 ONNX 模型进一步编译为 RKNN 文件，这一转换过程深入执行了图融合、内存复用以及算子重排等硬件级优化策略。基于此，模型能够充分释放 NPU 的异构计算潜能，为最终的端侧高效推理提供即插即用的执行文件。

3.3. 模型量化部署

针对 RK3588 开发板上的轻量化 U 型网络的部署和推理过程，本文按照如下流程对其进行量化部署(见图 4)：首先对原始 U 型网络进行轻量化设计，得到 MALUNet 模型，并进行归一化蒸馏，一次性层剪枝获

得轻量化网络；随后利用 RKNN-Toolkit2 生成 RKNN 模型文件；最后将生成的 RKNN 模型部署至 RK3588 开发板，实时输出分割结果，从而在保证分割精度的前提下实现了算法在嵌入式平台上的高性能运行。

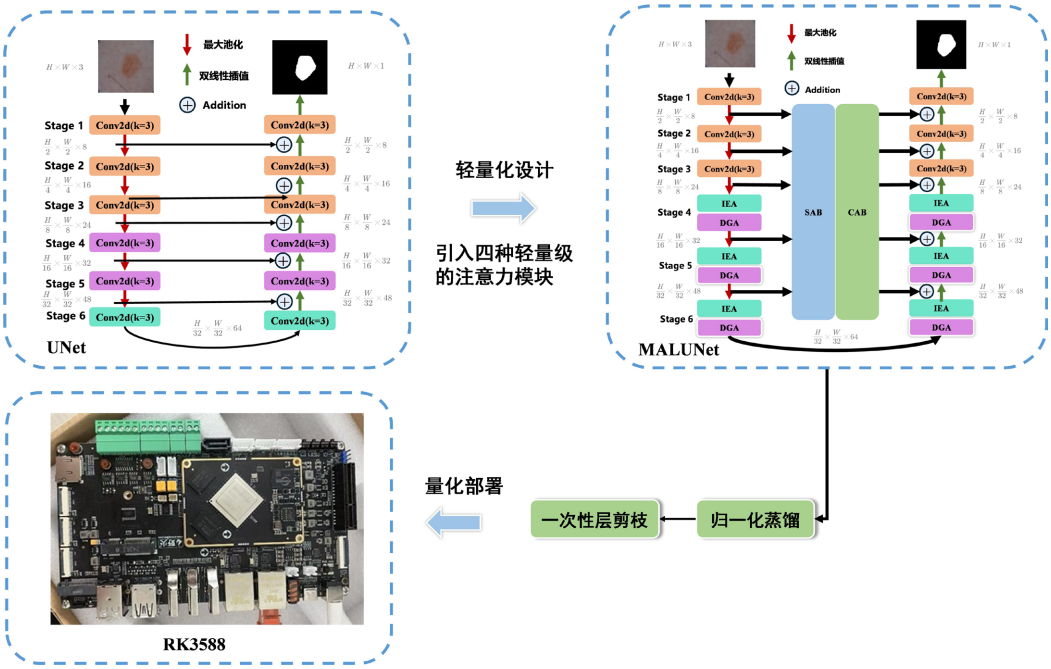

Figure 4. Deployment workflow of the image segmentation model
图 4. 图像分割模型部署流程图

4. 实验设置

4.1. 数据集

本实验选用 ISIC2017 (International Skin Imaging Collaboration 2017)皮肤病变数据集作为模型训练与评估的数据集。ISIC2017 数据集共包含 2150 张皮肤镜图像，每张图像均配有对应的像素级分割掩码标签，可用于监督式图像分割任务。图像内容涵盖多种皮肤病变类型，具有病变形态多样、边界复杂等特点，对模型的分割能力提出了较高要求。实验中，按照 7:3 的比例对数据集进行划分，其中训练集包含 1500 张图像，测试集包含 650 张图像。

4.2. 实验细节

MALUNet 网络的训练在单张 NVIDIARTX A4000 GPU 上完成。所有图像进行归一化并调整尺寸至 256 × 256，同时采用包括随机翻转、颜色变换和添加噪声等数据增强策略。损失函数采用 BCEDiceLoss，其表达式如公式(7)所示，学习率调度器采用 CosineAnnealingLR，优化器使用 AdamW，初始学习率为 1e−3，最大迭代次数为 300，最低学习率为 1e−5。训练周期设置为 300，批次大小为 16。

$$
\begin{aligned}
L_{BCE} &= -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(p_i\right)+\left(1-y_i\right)\log\left(1-p_i\right)\right], \\
L_{Dice} &= 1-\frac{2\sum_i P_i G_i}{\sum_i \left(P_i^2+G_i\right)}, \\
L_{BCEDice} &= \alpha \cdot L_{BCE}+\left(1-\alpha\right)\cdot\left(1-L_{Dice}\right).
\end{aligned}
\tag{7}
$$

其中，α 是一个超参数，用于调节 L_{BCE} 和 $L_{BCEDice}$ 之间的权重。

对于 MALUNet 教师模型的归一化蒸馏，所有学生模型的训练均在单张 NVIDIA GeForce RTX4060 GPU 完成。损失函数采用复合蒸馏损失，其表达式如公式(8)所示，学习率调度器采用 CosineAnnealingLR，优化器使用 AdamW，初始学习率为 1e−4，最大迭代次数为 50，最低学习率为 1e−6。蒸馏训练周期设置为 50，批次大小为 8，温度参数 $\tau = 3.0$，输出蒸馏权重 $\alpha = 0.5$，特征蒸馏权重 $\beta = 0.5$。

$$\begin{aligned} L &= L_{BCEDice} + \alpha \cdot L_{KD} + \beta \cdot L_{Feat}^{norm}, \\ L_{KD} &= \mathrm{KL}\left(\sigma\left(z_s/\tau\right) \| \sigma\left(z_t/\tau\right)\right) \times \tau^2, \end{aligned} \tag{8}$$

其中，$L_{BCEDice}$ 为原始分割任务的损失函数，表达式如公式(7)所示；L_{KD} 为输出蒸馏损失，即通过 KL 散度计算教师模型和学生模型输出之间的差异；L_{Feat}^{norm} 为归一化特征蒸馏损失，表达式如公式(5)所示。

MALUNet 的八种学生模型的单样本推理全部在 Rockchip RK3588 SoC 的 NPU 上完成，采用与训练相同图像预处理和后处理方法，采用 ISIC2017 训练集的前 150 个样本构成量化校准数据集。

4.3. 学生模型的设计

MALUNet 的通道数配置为[8, 16, 32, 64, 128]，总计参数量为 177,943，本文以 MALUNet 为基础，设计了八种不同轻量化程度的学生模型，具体设计的八种学生模型变体及其详细配置详见下表 1。

Table 1. System resulting data of standard experiment
表 1. 标准试验系统结果数据

模型名称	配置细节	参数量	参数压缩比
MALUNetSmall	各层通道数设置为[4, 8, 12, 16, 24, 32]	36,765	4.84×
MALUNetMedium	各层通道数配置为[8, 16, 20, 28, 40, 56]	132,991	1.34×
MALUNetShallow	将编码器与解码器构建为五层结构	109,991	1.62×
MALUNetNarrow	保持深度但减少每层复杂度	63,133	2.82×
MALUNetEfficient	全面采用深度可分离卷积以降低计算成本	18,803	9.46×
MALUNetAttLight	全面简化注意力机制模块以降低计算成本	126,092	1.41×
MALUNetGlobalAtt	使用全局注意力机制代替原有的多尺度注意力维度	126,718	1.40×
MALUNetMobile	结合 MobileNetV3 的逆残差结构与注意力机制进行重构	118,257	1.50×

上述八种变体分别代表了不同的轻量化优化方向，构成了完整的实验对照组。其中，MALUNetSmall 与 MALUNetMedium 主要通过调整通道数量来直接压缩参数量，旨在考察模型宽度对特征提取能力的影响；MALUNetShallow 通过减少编码器与解码器的层数来降低计算深度；MALUNetNarrow 标准卷积替换复杂注意力模块、使用 BatchNorm 替代 GroupNorm，全面降低模型复杂度；MALUNetEfficient 全面采用深度可分离卷积以降低计算成；MALUNetMobile 引入了 MobileNetV3 的逆残差结构；MALUNetAttLight 与 MALUNetGlobalAtt 分别尝试了简化注意力桥接结构与引入全局注意力机制。本文通过后续实验对比这八种模型在端侧的实际表现，将筛选出最适合该开发板部署环境的模型架构。

4.4. 评价指标

本文采用五类指标衡量分割与实时推理性能：平均交并比(mIoU)、Dice 系数、单样本推理时间(T_{inf})、单次推理浮点运算量(FLOPs)。

$$\mathrm{IoU}_i = \frac{TP_i}{TP_i + FP_i + FN_i}, \mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{IoU}_i, \tag{9}$$

$$\mathrm{Dice} = \frac{2\times TP_i}{2\times TP_i + FP_i + FN_i}, \tag{10}$$

其中，TP_i, FP_i, FN_i 分别表示第 i 类的真阳性、假阳性和假阴性，N 为类别总数。

5. 结果分析

Table 2. Distillation training losses comparison of student models on the training set
表 2. 学生模型在训练集上的蒸馏训练损失对比

模型名称	任务损失	输出蒸馏损失	归一化特征蒸馏损失	总损失
MALUNetSmall	0.3496	0.0007	0.5460	0.6229
MALUNetMedium	0.3348	0.0004	0.4287	0.5493
MALUNetShallow	0.3455	0.0004	0.4530	0.5722
MALUNetNarrow	0.3395	0.0005	0.5330	0.5632
MALUNetEfficient	0.3605	0.0008	0.7518	0.7368
MALUNetAttLight	0.3403	0.0004	0.4919	0.5865
MALUNetGlobalAtt	**0.3399**	**0.0004**	**0.3973**	**0.5388**
MALUNetMobile	0.4006	0.0020	1.3487	1.0760

Table 3. Distillation training losses comparison of student models on the validation set
表 3. 学生模型在验证集上的蒸馏训练损失对比

模型名称	任务损失	输出蒸馏损失	归一化特征蒸馏损失	总损失
MALUNetSmall	0.3827	0.0008	0.5172	0.6417
MALUNetMedium	0.3779	0.0004	0.4351	0.5959
MALUNetShallow	0.3827	0.0004	0.4335	0.5996
MALUNetNarrow	0.3915	0.0005	0.4535	0.6014
MALUNetEfficient	0.3911	0.0008	0.7364	0.7598
MALUNetAttLight	0.3814	0.0004	0.4590	0.6111
MALUNetGlobalAtt	**0.3815**	**0.0004**	**0.3775**	**0.5705**
MALUNetMobile	0.4287	0.0018	1.2255	1.0423

表 2、表 3 显示，MALUNetGlobalAtt 均获得了在 ISIC2017 数据集上的归一化蒸馏训练获得了最低的总损失值，接下来在我们将上述八种模型量化部署至 RK3588 开发板，并进行单样本推理，并与原始 MALUNet 模型在 PC 端的测试结果做对比。

经测试，原始 MALUNet 模型在单张 NVIDIA GeForce RTX4060 GPU 上的 mIoU 为 0.8847，Dice 系数为 0.9194，单样本推理时间为 0.2503 s，FLOPs 为 355886。

选取索引为 0 的样本图片进行单样本推理，结果见表 4，同时选取索引为 0，52，147 的样本图片进行单样本推理，结果见图 5。在同一张图片上，MALUNetGlobalAtt 模型取得了更高的 mIoU 和 Dice 系

数，具有更好的模型性能，并且推理时间大幅度降低。

Table 4. Comparison of experimental results of student models
表 4. 学生模型实验结果对比表

模型名称	mIoU	Dice 系数	T_{inf} (s)	推理时间降低率	FLOPs
MALUNetSmall	0.8122	0.8963	0.1854	25.9%	73,530
MALUNetMedium	0.6114	0.7005	0.0588	76.5%	265,982
MALUNetShallow	0.7154	0.8341	0.0131	94.8%	428,978
MALUNetNarrow	0.6559	0.7922	0.0100	96.0%	126,266
MALUNetEfficient	0.7594	0.8632	0.0140	94.4%	37,606
MALUNetAttLight	0.7640	0.8662	0.0070	97.2%	252,184
MALUNetGlobalAtt	**0.8126**	**0.8966**	**0.0100**	**96.0%**	**253,436**
MALUNetMobile	0.8012	0.8896	0.0201	92.0%	236,514

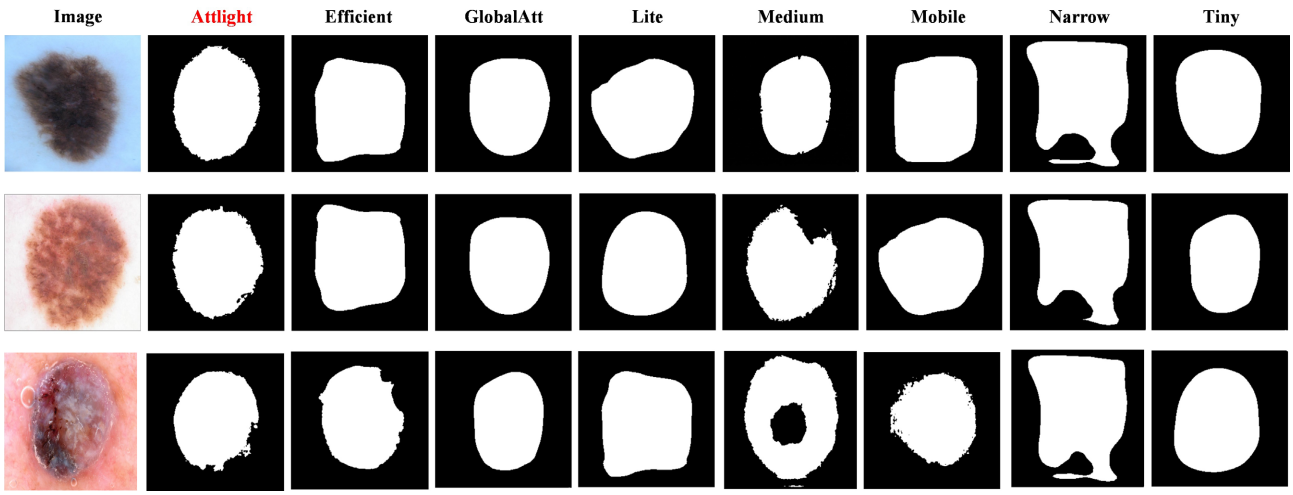

Figure 5. Comparison of experimental results of student models
图 5. 学生模型实验结果对比图

6. 结论

本文针对医学影像分割模型在边缘设备上难以兼顾精度与实时性的问题，设计了轻量化网络MALUNet，并基于国产瑞芯微 RK3588 平台完成了全流程量化部署。研究通过引入空洞门控、反转外部注意力及特征桥接模块优化网络结构，并结合归一化知识蒸馏与一次性层剪枝策略，有效解决了复杂模型在受限算力下的运行瓶颈。在 ISIC2017 数据集上的实验结果表明，本文提出的轻量化 MALUNetGlobalAtt 综合性能最优，在保持较高分割精度(mIoU 0.8126)的同时，相比原始模型推理时间降低了 96%，成功实现了在 RK3588 NPU 上的毫秒级高效运行。本文提出的软硬协同优化方案验证了深度医学影像算法在国产边缘计算平台落地的工程可行性，为智慧医疗的端侧应用提供了具有实用价值的技术参考。

致　谢

在论文完成后，首先，我们非常想要感谢指导老师的辛勤指导。在我们撰写论文的过程中，从一开

始论文的选题，再到论文的构思与撰写方式，我们得到了指导老师悉心细致的教诲和无私的帮助，他那广博的知识、深厚的学术素养、严谨的治学精神使我们这一生都会终身得到受益。在此，我们向指导老师表示真挚的感谢。

在论文的写作过程中，我们也得到了许多同学的宝贵建议，也在网上看到了诠释了开源精神的先辈们，若没有他们将辛苦搜集整理出来的数据开源，我们将寸步难行，在此，对他们表示真诚的谢意。

最后的最后，感谢所有关心、支持、帮助过我们的良师益友。

参考文献

[1] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W. and Frangi, A., Eds., *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2015, Springer, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[2] Du, G., Cao, X., Liang, J., Chen, X. and Zhan, Y. (2020) Medical Image Segmentation Based on U-Net: A Review. *Journal of Imaging Science and Technology*, **64**, 020508-1-020508-12. https://doi.org/10.2352/j.imagingsci.technol.2020.64.2.020508

[3] Cui, K. and Tian, Q.C. (2024) Review of Medical Image Segmentation Algorithms Based on U-Net Variants. *Journal of Computer Engineering & Applications*, **60**, 32.

[4] Zhang, R. and Chung, A.C.S. (2024) EfficientQ: An Efficient and Accurate Post-Training Neural Network Quantization Method for Medical Image Segmentation. *Medical Image Analysis*, **97**, Article ID: 103277. https://doi.org/10.1016/j.media.2024.103277

[5] Liu, Q., Zhou, S. and Lai, J. (2023) EdgeMedNet: Lightweight and Accurate U-Net for Implementing Efficient Medical Image Segmentation on Edge Devices. *IEEE Transactions on Circuits and Systems II*: *Express Briefs*, **70**, 4329-4333. https://doi.org/10.1109/tcsii.2023.3291168

[6] Ramesh, K.K.D., Kumar, G.K., Swapna, K., Datta, D. and Rajest, S.S. (2021) A Review of Medical Image Segmentation Algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, **7**, e6. https://doi.org/10.4108/eai.12-4-2021.169184

[7] Bu, D., Sun, B., Sun, X. and Guo, R. (2024) Research on YOLOv8 UAV Ground Target Detection Based on RK3588. 2024 2*nd International Conference on Computer*, *Vision and Intelligent Technology* (*ICCVIT*), Huaibei, 24-27 November 2024, 1-5. https://doi.org/10.1109/iccvit63928.2024.10872495

[8] Ruan, J., Xiang, S., Xie, M., Liu, T. and Fu, Y. (2022) MALUNet: A Multi-Attention and Light-Weight UNet for Skin Lesion Segmentation. 2022 *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), Las Vegas, 6-8 December 2022, 1150-1156. https://doi.org/10.1109/bibm55620.2022.9995040

[9] Guo, M., Liu, Z., Mu, T. and Hu, S. (2022) Beyond Self-Attention: External Attention Using Two Linear Layers for Visual Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 5436-5447. https://doi.org/10.1109/tpami.2022.3211006

[10] LeCun, Y., Denker, J. and Solla, S. (1989) Optimal Brain Damage. *Advances in Neural Information Processing Systems*, **2**, 1-5.

[11] Hassibi, B., Stork, D.G. and Wolff, G.J. (1993) Optimal Brain Surgeon and General Network Pruning. *IEEE International Conference on Neural Networks*, San Francisco, 28 March-1 April 1993, 293-299. https://doi.org/10.1109/icnn.1993.298572

[12] Han, S., Mao, H. and Dally, W.J. (2015) Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. arXiv: 1510.00149.

[13] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. arXiv: 1503.02531.

[14] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C. and Bengio, Y. (2014) FitNets: Hints for Thin Deep Nets. arXiv: 1412.6550.

[15] Courbariaux, M., Hubara, I., Soudry, D., *et al.* (2016) Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to+ 1 or −1. arXiv: 1602.02830.

[16] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., *et al.* (2018) Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 2704-2713. https://doi.org/10.1109/cvpr.2018.00286

[17] Zhang, D., Li, S., Chen, C., *et al.* (2024) LAPTOP-Diff: Layer Pruning and Normalized Distillation for Compressing

Diffusion Models. arXiv: 2404.11098.

[18] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M.W. and Keutzer, K. (2022) A Survey of Quantization Methods for Efficient Neural Network Inference. In: Thiruvathukal, G.K., Lu, Y.H., Kim, J., Chen, Y.R. and Chen, B., Eds., *Low-Power Computer Vision*, Chapman and Hall/CRC, 291-326. https://doi.org/10.1201/9781003162810-13